Towards Real-Time Detection of Squamous Pre-Cancers from Oesophageal Endoscopic Videos

Xiaohong Gao
Department of Computer Science
Middlesex University
London, United Kingdom
x.gao@mdx.ac.uk

Barbara Braden
Gastroenterologist Translational
Gastroenterology Unit
John Radcliffe Hospital
Oxford, United Kingdom
braden@em.uni-frankfurt.de

Stephen Taylor
MRC WIMM Centre of Computational Biology
MRC Weatherall Institute of Molecular Medicine
Oxford, United Kingdom
stephen.taylor@imm.ox.ac.uk

Wei Pang
Department of Computing Science
University of Aberdeen
Old Aberdeen, United Kingdom
pang.wei@abdn.ac.uk

***Abstract*— This study investigates the feasibility of applying state of the art deep learning techniques to detect precancerous stages of squamous cell carcinoma (SCC) cancer in real time to address the challenges while diagnosing SCC with subtle appearance changes as well as video processing speed. Two deep learning models are implemented, which are to determine artefact of video frames and to detect, segment and classify those no-artefact frames respectively. For detection of SCC, both mask-RCNN and YOLOv3 architectures are implemented. In addition, in order to ascertain one bounding box being detected for one region of interest instead of multiple duplicated boxes, a faster non-maxima suppression technique (NMS) is applied on top of predictions. As a result, this developed system can process videos at 16-20 frames per second. Three classes are classified, which are 'suspicious', 'high grade' and 'cancer' of SCC. With the resolution of 1920x1080 pixels of videos, the average processing time while apply YOLOv3 is in the range of 0.064-0.101 seconds per frame, i.e. 10-15 frames per second, while running under Windows 10 operating system with 1 GPU (GeForce GTX 1060). The averaged accuracies for classification and detection are 85% and 74% respectively. Since YOLOv3 only provides bounding boxes, to delineate lesioned regions, mask-RCNN is also evaluated. While better detection result is achieved with 77% accuracy, the classification accuracy is similar to that by YOLOYv3 with 84%. However, the processing speed is more than 10 times slower with an average of 1.2 second per frame due to creation of masks. The accuracy of segmentation by mask-RCNN is 63%. These results are based on the date sets of 350 images. Further improvement is hence in need in the future by collecting, annotating or augmenting more datasets. (*Abstract*)**

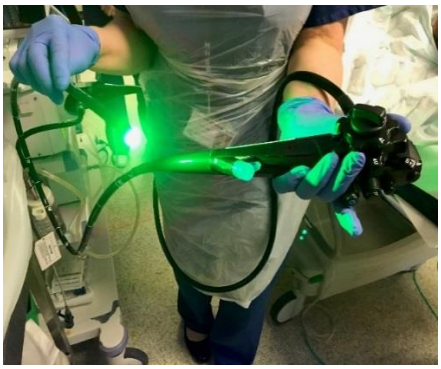
Keywords*— *oesophagus endoscopy, pre-cancer detection, deep learning, segmentation, real-time video processing (key words)

I. Introduction

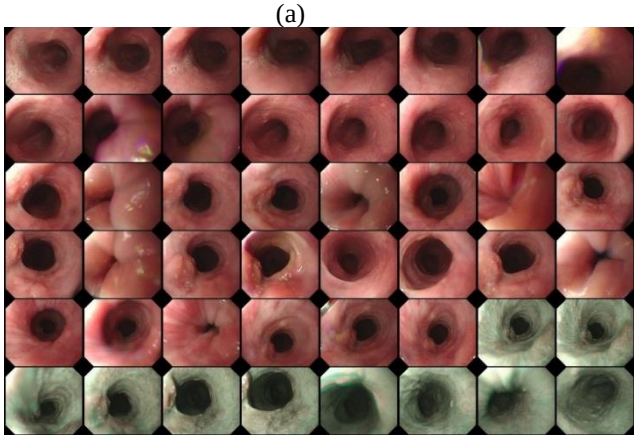
Oesophagus cancer (OC), or cancer of the gullet, is the 8th most common cancer worldwide [1] and the 6th leading cause of cancer-related death [2]. Two main histological types represent the most majority of all oesophageal cancers, which are adenocarcinoma and squamous cell carcinoma cancer (SCC). Worldwide, about 87% of all oesophageal cancers are SCC with the highest incidence rates occurring in Asia, the Middle East and Africa [3,4].

While the five-year survival rate of oesophagus cancer is less than 20% [5], the rate can be improved significantly to more than 90% if the cancer is detected in its early stages when it still can be treated endoscopically [6]. Hence there is a clinical urgency to improve the detection of oesophageal pre-cancerous stages, e.g. dysplasia, to allow endoscopic treatment and monitoring of affected patients.

The procedure to perform oesophageal inspection is to insert an endoscopic camera as illustrated in Figure 1 (a), into the food pipe of the patient in concern, whereby the appearance inside the oesophageal tube can be visualised through a computer monitor connected to the camera (Figure 1 (b)).



(a)



(b)

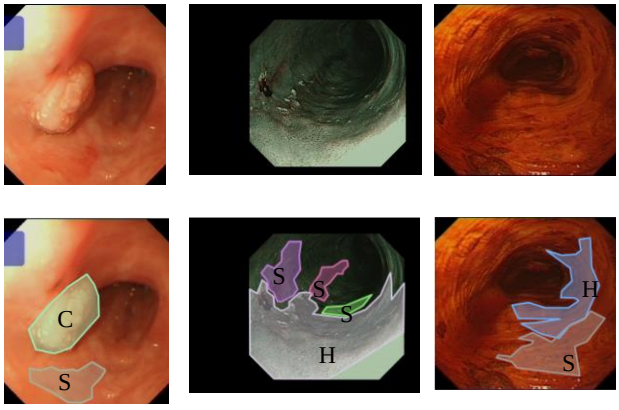
Figure 1. (a) The oesophagus camera. (b) A montage display of a clip of an endoscopic video where the last row showing the narrow-band imaging (NBI), whereas the top rows depicting conventional white light endoscopy (WLE).

As it can be seen from Figure 1, due to the limitation of manoeuvring space of the endoscopic camera, the acquired images depict not only the surface of oesophageal tube walls but also artefacts (e.g. colour misalignment, burry, secularity, saturation, bubbles, saturation, etc.).

A. *Challenges for detecting oesophageal squamous cancer*

Precancerous stages (dysplasia in the oesophageal squamous epithelium) and early stages of SCC are easily missed at the time of conventional white light endoscopy (WLE) as these lesions grow usually flat with only subtle changes in colour and in microvasculature as demonstrated in Figure 2 for those suspicious regions ('S' and 'H'), where 'C' refers to 'cancer', 'H' for 'High grade' of possible of cancer and 'S' for 'suspicious'. To overcome this shortcoming while viewing WLE images, narrow-band imaging (NBI) can be turned on to display only two wavelengths (415 nm (blue) and 540 nm (green)) (Figure 2(b)) to improve the visibility of those suspected lesions by filtering out the rest of colour bands. Another approach is dye-based chromoendoscopy, i.e. Lugol's staining technique, which highlights dysplastic abnormalities by spraying iodine [7] (Figure 2(c)).

While NBI technique improves the visibility of the vascular network and surface structure, it mainly facilitates the detection of unique vascular and pit pattern morphology that are present in neoplastic lesions [8], whereas precancerous stages can take a variety of forms. For Lugol's staining approach, many patients react uncomfortably to the spray.



(a) (b) (c)

Figure 2. Examples of SCC where C=cancer, S=suspicious, H=High grade. Top row: original images; bottom: with masks. (a) WLE; (b) NBI; (c) Lugol's.

It is therefore of clinical priority to have a computer assisted system to help clinicians to detect and highlight those potential suspected regions for further examinations. While there exists a number of promising results for computer-aided recognition of early neoplastic oesophageal lesions from endoscopic still images [9, 10], there are less algorithms that are applicable to real-time endoscopy to allow computer-aided decision-making during endoscopy at the point of examination. Furthermore, the existing studies focus mainly on the classification of endoscopic images between normal and abnormal stages with little work providing bounding boxes of the suspicious regions (detection) and delineating (segmentation).

B. *The State Of the Art Deep Learning Models for Object Detection and Segmentation*

With the advances of deep learning techniques, accurate detection of objects has been realised by many neural network models. One group of the latest ones would be the region-based Convolutional Neural Networks, or R-CNN [11]. This family of R-CNN comprises two major steps, e.g. (1) extracting region proposals (i.e., regions of an image that potentially contain objects) using an algorithm such as *Selective Search* [12], and (2) obtaining CNN features from each region independently for classification, by applying a classifier, e.g. *Support Vector Machine* (SVM).

While R-CNN is robust with discriminative features learned by the CNN, it tends to be very slow. As a result, fast R-CNN algorithm [13] is developed. While still utilising *Selective Search* to obtain region proposals, fast-RCNN introduces a novel concept, which is *Region of Interest (ROI) Pooling module* by extracting a fixed-size window from the feature map and employing these features to obtain the final class label and bounding box.

While this fast network performs end-to-end training, its performance suffered dramatically at inference (i.e., prediction) by being dependent on *selective search*. To make the R-CNN architecture even *faster*, region proposals need to be incorporated *directly* into the R-CNN, which leads to the generation of *Faster R-CNN* [14] by introducing the *Region Proposal Network (RPN)* that takes region proposal *directly* into the architecture, alleviating the need for the application of selective search algorithm, which significantly increases the detection/segmentation rates, being able to process 3 to 10 frames per second for COCO datasets (i.e. people, cars, boats, etc.) depending on the size of an image (a typical image with a resolution of 640 ×480 pixels).

C. *Real-Time Processing of Videos*

While faster R-CNN can process images quickly, there is still a long way to go to perform real-time detection, especially when image sizes are varying and large (e.g. 1920×1080 in this study). Since R-CNN family performs detection takes place in two steps, further improvement takes place to combine two steps into one. One of such models is YOLO family, referring to 'you only look once' [15].

Since the potential bounding box candidates can be infinite, to ensure the speedy process, YOLO approach skips this region proposal stage and runs detection directly over a dense sampling of possible locations. As a one-stage object detector, YOLO is able to achieve fast processing speed. However, the accuracy rate for recognizing irregularly shaped objects or a group of small objects is not as good as R-CNN models, which is due to a limited number of bounding box candidates to work on. Hence YOLO version 2, YOLOv2 [16] has been proposed to improve prediction accuracy while maintaining the processing speed by introducing batch normalisation, fine-tuning base model image resolution, convolutional layer for anchor box

detection, k-mean clustering of box dimensions, and direct location prediction. More recently, YOLOv3 [17] is developed by applying a number of design techniques on YOLOv2, again attempting to improve both performance accuracy and processing speed.

Other one-stage models include *Single Shot Detector* (SSD) [18] that attempts to apply convolutional neural network's pyramidal feature hierarchy for efficient detection of objects of various sizes, and RetinaNet [19], a one-stage dense object detector. For RetinaNet, there are two crucial building blocks, which are *featurised image pyramid* and the use of *focal loss*. It has been found that overall YOLOv3 performs better and faster than SSD. Although YOLOv3 does not perform as well as RetinaNet but appears to be 3.8 times faster [19].

While those networks (YOLO or R-CNN family) can be improved to perform fast, they only generate four sets of *(x, y)* coordinates representing the *bounding box* of an object in an image. The bounding box itself does not provide information on which pixels belong to the foreground object and which pixels belong to the background. In particular, in the case of the detection of suspected SCC regions, biopsy needs to be collected from the right diseased spot. Hence, segmentation while detection simultaneously might also be in need.

In this study, both YOLOv3 and mask-RCNN [20] are evaluated using endoscopic data, aiming at real-time detection of pre-squamous-cancer from video images.

II. Methods

Figure 3 depicts the flowchart of the system developed in this study, including two machine learning models, one for determination of artefact and another for detection, classification and segmentation.

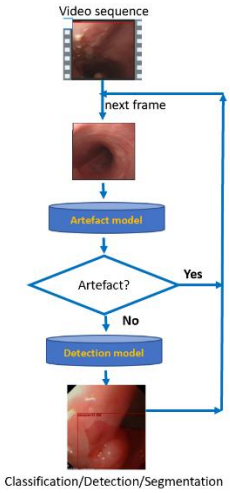


Figure 3. The flow chart of the developed system for processing oesophagus videos.

For classification of artefact, conventional deep learning convolutional neural network (CNN) is applied, whereas for detection, two state of the art deep learning architectures are implemented and evaluated, which are mask-RCNN and Yolov3 models.

A. Datasets

At present, 352 still images extracted from 15 subjects' endoscopic videos with SCC have been collected from Oxford NHS University Hospital, UK. There videos last from 10 to 30 minutes with 50 frames per second (FPS). The resolution of these videos is 1920 × 1080 pixels whereas still images have varying sizes between 256 × 256 and 1920 × 1080. Although in terms of frame numbers, there are over 1 million still images. Many of those frames/images are of normal or with artefact. As a result, at present, around 350 images are collected and annotated by the clinician in the team labelled using the software of VGG Annotation Annotator (VIA) [25]. Images are labelled with three classes ('high-grade', 'suspicious', 'cancer') of SCC as exemplified in Figure 2, leading to 537 masks. These images are composed of white light endoscopic (WLE) images, narrow band images (NBI) and Lugol's. Since the training is built upon transfer learning with backbone models of Resnet-101-RPN for mask-RCNN and Darket-53 for YOLOv3, this group of images appears to be sufficient for evaluation. The ratio between training and evaluation is set to be 0.9 to 0.1, which is split randomly.

B. Detection of Misalignment of Colour Channels

For a clip of videos acquired in real time, nearly half of the frames contain artefact due to the motion of video cameras navigating within confined space as demonstrated in Figure 5.

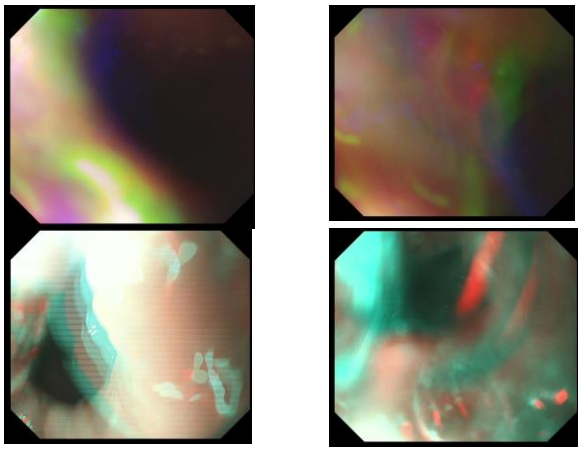


Figure 4. Examples of video frames with artefact. Top row: WLE. Bottom row: NBI.

While artefact can have many forms, including saturation, blurring, specularity as well as instrument, at present, this study only focuses on colour channel misalignment including blurry. During the acquisition of endoscopic videos, colour channel images (e.g. RGB) are commonly acquired sequentially at different time instances for each location to be filmed and merged into the final video frame of that location. However, when the speed of the camera is too fast, the acquisition colour channel instances of an image might correspond to two or more different locations, e.g. the location in R channel is different to that in B or G channel. As a result, these channels are misaligned in the resulting

video frames with the appearances of being unnatural, highly colourful and stroboscopic (Figure 4). To avoid unaccurate diagnosis as well as to speed up the process of videos, these frames are determined first and ignored for further detection of SCC.

In this study, CIELAB colour space is applied instead of RGB to simulate human colour perception. In doing so, an image is firstly converted from RGB to CIELAB or LAB images by using OpenCV library. Then conventional CNN model, e.g. AlexNet, is applied to train a binary classifier for determine frames with artefact or not. With 800 images for each class for training and 200 each for evaluation and testing, the classification accurate is 96%.

This pre-processing step takes about 0.017 second per frame (SPF).

C. Mask-RCNN

In this study, Mask R-CNN [20] with Tensorflow and Keras libraries (for speeding process) is implemented and evaluated. Figure 5 depicts the work flow of mask R-CNN network that is applied in this study.

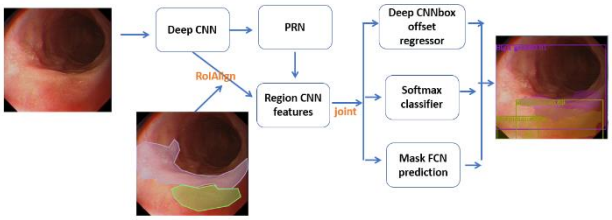


Figure 5. Illustration of workflow of mask R-CNN network.

Based on the framework of Faster R-CNN, Mask R-CNN added a third branch for predicting an object mask in parallel with the existing branches for classification and localization. The mask branch is a small fully-connected network applied to each region of interest (RoI), predicting a segmentation mask in a pixel-to-pixel manner through leveraging a Region Proposal Network (RPN).

As a result, the multi-task loss function ($\mathcal{L}$) is formulated in Eq. (1) by combining the loss of classification, localization and segmentation mask:

$$\mathcal{L} = \mathcal{L}_{class} + \mathcal{L}_{box} + \mathcal{L}_{mask} \quad (1)$$

Where

$$\mathcal{L}_{mask} = \frac{1}{m^2}\sum_{1\le i,j\ge m}[y_{ij}\log\hat{y}_{ij}^{k} + (1-y_{ij})\log(1-\hat{y}_{ij}^{k})] \quad (2)$$

In Eq. (2), y_{ij} is the label of a cell (i,j) in the true mask for the region of size of $m \times m$; $\hat{y}_{ij}^{k}$ is the predicted value of the same cell in the mask learned for the ground-truth class k.

$$\mathcal{L}_{class}(p,k) = -\log p_k \quad (3)$$

$$\mathcal{L}_{box}(t^k, g) = \sum_{i\in\{x,y,w.h\}} L_1^{smooth}(t_i^k - g_i) \quad (4)$$

where

$$L_1^{smooth}(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \quad (5)$$

And p the probability of class k; t^k refers to predicted bounding box correction and g the true bounding box.

D. Yolov3

Figure 6 demonstrates the workflow of YOLOv3 applied in this study. The network takes input of images with ground truth bounding boxes and then predict feature maps in the form of 3D tensors, which are corresponding to three scales (i.e. 13×13, 26×26, 52×52) designed to detect objects in varying sizes. For example, for the scale of 13×13, the input image is divided into 13×13 grid cells, whereby each cell corresponds to a voxel of size of 1×1×24 inside a 3D tensor. For each grid cell, the network provides predictions of 3 prior boxes of different sizes to choose from whereby the box that overlaps ground truth bounding box the most will be selected to be the final box. In this study, there are 3 classes (i.e. 'suspicious', 'high grade' and 'cancer') plus a background (1+3) whereas a bounding box is represented using 4 corner coordinates for each of 3 boxes. Hence 24 refers to $24=3_{\text{predicted-box}} \times (4_{\text{coordinate}}+1_{\text{bg}}+3_{\text{class}})$.

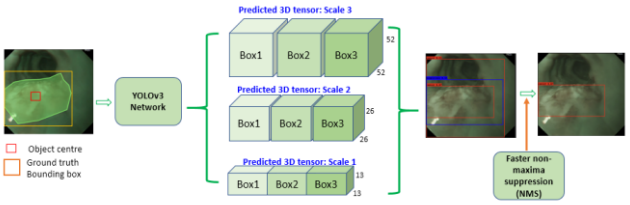


Figure 6. The workflow of YOLOv3 applied this this study.

In order to choose the initial size of those 3 prior boxes, K-mean clustering is utilised to classify the total bounding boxes (bboxes) from SCC dataset into 9 clusters before training, resulting in 9 sizes chosen from 9 clusters, 3 for each of three scales, which are [(33,27), (57,71), (73,37), (116,77), (117,128), (202,249), (210,152), (316,249), (377,362)] for this collection of oesophagus videos. This prior information potentially expedites the learning process of the network and computes box offset/coordinate more precisely. It has a backbone of Darknet-53 containing 53 convolutional layers. The loss function is expressed in Eq. (6), penalising on the objectness score prediction for bounding boxes (Eq.(7)) responsible for predicting objects (ideally being 1), the bounding boxes having no objects (Eq. (8)) (ideally being 0), and the class prediction (Eq.(9)) for the bounding box which predicts the objects.

$$L = L_{\text{bbox-obj}} + L_{\text{no-obj}} + L_{\text{class-prediction}} \quad (6)$$

Where

$$\mathcal{L}_{bbox-obj} = \lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \quad (7)$$

$$L_{\text{no-obj}} = \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \quad (8)$$

$$\mathcal{L}_{class-prediction} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} (C_i - \hat{C}_i)^2 + \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} (C_i - \hat{C}_i)^2 + \sum_{i=0}^{S^2} I_i^{obj} \sum_{c \epsilon classes}^{B} (p_i(c) - \hat{p}_i(c))^2 \quad (9)$$

where λ_{coord} denotes the weight on a bounding box with centre coordinates of (x, y) and width (w) and height (h). C refers to class and p the probability. I_{ij}^{obj} indicates the weights to be learned in the network.

The implementation of both YOLOv3 and Mask-RCNN is conducted by applying Keras and Tensorflow libraries in Python language as given in [21, 22] under Windows 10 systems with one GPU (Nvidia GeForce GTX 1060) and 16 GB memory. Similar to training on COCO datasets [23], for mask-RCNN, during the prediction stage, each of the 300 ROIs go through non-maxima suppression [24] and the top 100 detection boxes are kept, resulting in a 4D tensor of *100 x 3 × 15 × 15* where *3* is the number of class labels (i.e. 'suspicious', 'High-grade' and 'cancer') in the dataset and *15 × 15* refers to the size of each of the *3* masks. The learning rate is 0.0001.

For mask-RCNN, the object classifier is built upon HoG coupled with linear SVM whereas for YOLOv3, logistic regression is being orchestrated. All these approaches inevitably detect multiple bounding boxes with varying classes surrounding an object in an image. While thresholding can be applied to leverage this phenomenon to a certain degree, too high threshold can remove all the bboxes altogether. Hence a faster non-maxima suppression technique [24] (NMS) is applied to compute the overlap ratios and determine which bounding boxes to be ignored leading to the final detection results. In this study, the overlapping threshold is set to be 0.8 to limit the duplications of near identical boxes labelled with different classes.

III. Results

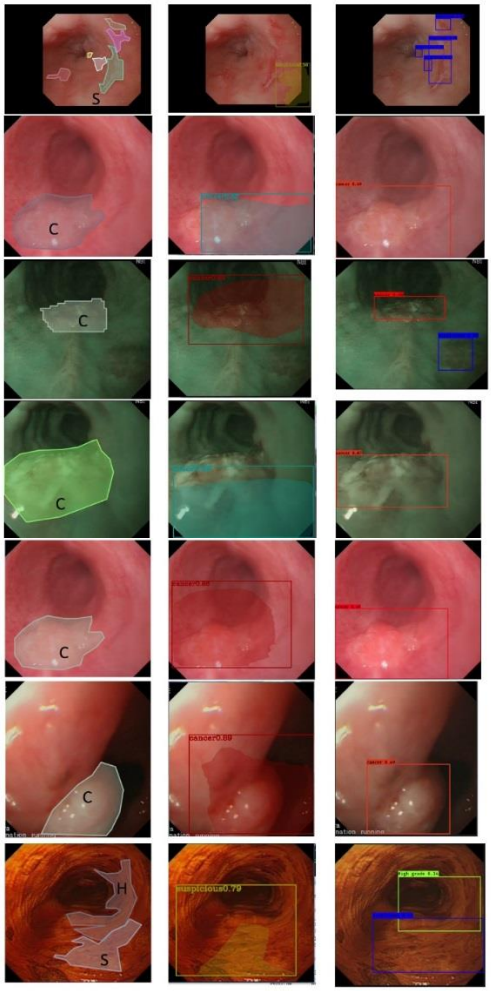
Table 1 illustrates the classification, segmentation and detection results. The training takes 160 epochs for Mask-RCNN and 100 epochs for YOLOv3 when learning rate of 0.99e-7 is reached. The evaluating results are from 35 examples, 10% of the total data whereas the rest were used for training.

TABLE 1. EVALUATION RESULTS FOR OESOPHAGEAL SCC IMAGES USING BOTH YOLOV3 AND MASK-RCNN APPROACHES.

	Class-loss	**BBOX-loss (mAP)**	**mask-loss (mAP)**	**Testing time (SPF)**
Yolov3-Keras (threshold=0.1, IOU=0.3)	**0.1495**	0.2557		**0.064 - 0.101**
Mask-RCNN (threshold=0. 50)	0.1575	**0.2296**	0.3616	0.450 - 1.20

Both networks of mask-RCNN and Yolov3 perform well when it comes to classification with 84% and 85% accuracy respectively with YOLOv3 performs slightly better. For detection of bounding boxes, mask-RCNN performs better with loss of 22.96% in comparison with YOLOv3 with the loss of 25.57%. In addition, mask-RCNN provides masks for detected regions with a loss of 36.16%. However, when it comes to processing time, YOLOv3 runs more than 10 times faster with 10 frames or more per second, leading to real-time processing if videos are of 8 frames per second (FPS), which fits the perception of human eyes.

Figure 7 demonstrates a number of detection and segmentation results for images of WLE, NBI and Lugol's using both YOLOv3 and mask-RCNN. For YOLOv3, only bounding boxes are produced. With regard to mask-RCNN, yellow, red and cyan colours refer to classes of 'High grade', 'suspicious' and 'cancer' whereas green, blue and red for YOLOv3 approach. The left most column refers to ground truth where C = 'cancer', H = 'High grade' and S = 'suspicious'.



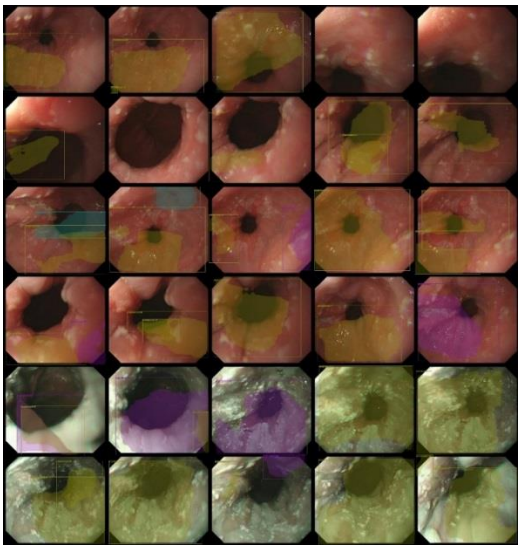
(a) (b) (c)

Figure 7. Illustration of training results using mask-RCNN (middle column) and YOLOv3 (right column). Left column: ground truth where S= suspicious, H=High grade, C=cancer.

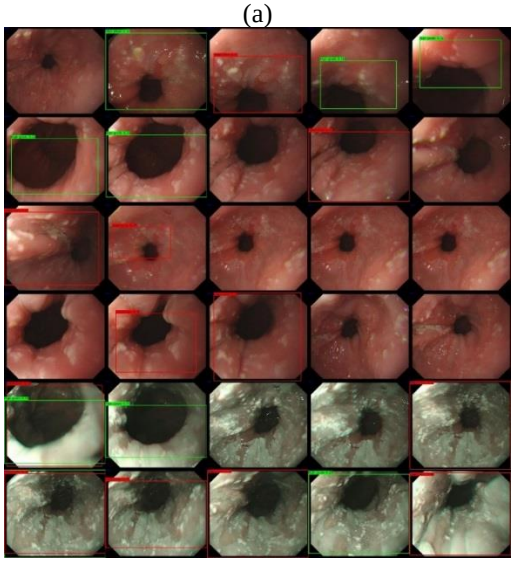
In this study, duplicated boxes for one single region is dealt with applying NMS, when if more than 80% of two boxes are overlapping as depicted in Figure 8 for YOLOv3 approach.

Figure 8. Illustration of removing duplicated boxes labelled with different classes using NMS approach when overlapping exceeding 80%. The example showing the result from YOLOv3.

Figure 9 displays a montage of a clip of video that has SCC for both mask-RCNN (Figure 8(a)) and YOLOv3 (Figure 8(b)). Each frame has a resolution of 1201×1051 pixels. The processing time for each frame is around 1.2 SPF for mask-RCNN and 0.08 SPF for YOLOv3. While the bounding boxes are closer to the ground truth applying the mask-RCNN, the segmented masks appear to be mis-matched to a certain extend.



(a)



(b)

Figure 9. Result of processing a video clip for both mask-RCNN (a) and YOLOv3 (b) approaches.

IV. DISCUSSION AND CONCLUSION

This work investigates the feasibility of achieving real-time processing of endoscopic video images by applying deep learning architectures of mask-RCNN and YOLOv3 to detect and segment oesophageal endoscopic images in an aim to determine precancer status of squamous cell cancer (SCC). Three classes are studied, which are 'cancer', 'high grade' and 'suspicious' of SCC. With regard to classification, 85% and 84% accuracy is achieved for mask-RCNN and YOLOv3 respectively whereas 74% and 77% for detection. With regard to segmentation, only mask-RCNN can produce masks and achieves 68% accuracy. When it comes to speed while testing videos, about 1.2 second per frame (SPF) is realised using mask-RCNN, which is 10 times slower when applying YOLOv3 with 0.1 SPF.

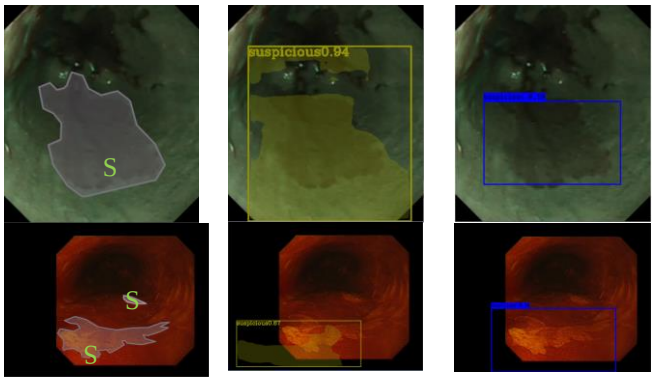
While training time can be expedited with the increasing number of GPUs for batch processing, i.e. 8 GPUs, during testing stage, only 1 GPU is in need as a video is processed frame by frame. Hence the quality of this single GPU plays a crucial role in real-time detection. In our study, when running with NVIDIA Quatro K4000 GPU card, the processing speed is ~1s/per frame while applying YOLOv3. However, when applying GeForce GTX 1060 GPU card (with Cuda 9.2), the processing time is about 10 times faster at ~0.1s/per frame, i.e. at 8 frames/per second.

Although processing each frame takes 0.1s, when the system works on videos, many frames (~40%) are ignored after undertaking detection of colour misalignment, taking up an average of ~0.006 second, leading to the processing time for a clip videos at 16-20 frames/second (Figure 8).

In addition, determination of score thresholds for the final results should be performed when more training datasets are available. Although both mask-RCNN and YOLOv3 apply transfer learning, this dataset (N=352) for both training and evaluation appear to be small, especially, for mask-RCNN, where the segmented masks appear to be mismatched at many cases, as presented in Figure 10. This mis-masking turn to be more severe for NBI and Lugol's images. As a result, bounding boxes turn to be sufficient to locate diseased regions.

For mask-RCNN, in this study, the number of region candidates is set to be 300 whereas top tanked 100 are selected for further process, in order to accelerate the processing speed. However, the repeatability of detection and segmentation appear to be affected to a certain extent, e.g. a suspected region has a bbox for the first run and has no bbox the next time. Future investigation will be carried out to determine if increasing the region candidates (e.g. 1000) or training dataset will improve the reproduction.

For YOLOv3, three candidate boxes are selected for each cell. Hence the results are highly repeatable.



(a) (b) (c)

Figure 10. Mismatch of masks from mask-RCNN approach. (a) original images with ground truth mask; (b) mask-RCNN approach; (c) YOLOv3 approach. Top row: NBI images; Bottom: Lugol's images, where S='suspicious'.

In the future, more data will be collected, annotated or augmented to ensure both speedy and accurate detection and segmentation of precancerous regions of SCC. Another approach is to improve the implementation of algorithms, for example, using Pytorch as implemented in mask-RCNN benchmark, which is undergoing investigation currently.

With regard to detection of artefact of colour channel mis-alignment, CIECAM02 [26] colour appearance model with JCH (Lightness, Colourfulness, and Hue) space could be more accurate than CIELAB. However, it takes more than 2 seconds to process each frame, much longer than the detection of SCC. Further investigation will be carried out to optimise CIECAM02 algorithm.

ACKNOWLEDGMENT

This project (Endo.AI, 2019-2020) is funded by the Cancer Research UK (CRUK). Their financial support is gratefully acknowledged.

REFERENCES

[1]. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. American Cancer Society; 2018 Nov;68(6):394–424.

[2]. Pennathur A, Gibson MK, Jobe BA, Luketich JD. Oesophageal carcinoma. Lancet. 2013 Feb 2;381(9864):400–12.

[3]. Arnold M, Laversanne M, Brown LM, Devesa SS, Bray F. Predicting the Future Burden of Esophageal Cancer by Histological Subtype: International Trends in Incidence up to 2030. Am J Gastroenterol. Nature Publishing Group; 2017 Aug;112(8):1247–55.

[4]. Arnold M, Soerjomataram I, Ferlay J, Forman D. Global incidence of oesophageal cancer by histological subtype in 2012. Gut. BMJ Publishing Group; 2015 Mar;64(3):381–7.

[5]. Siegel R, Ma J, Zou Z, Jemal A. Cancer statistics, 2014. CA Cancer J Clin. 3rd ed. American Cancer Society; 2014 Jan;64(1):9–29.

[6]. Shimizu Y, Tsukagoshi H, Fujita M, Hosokawa M, Kato M, Asaka M. Long-term outcome after endoscopic mucosal resection in patients with esophageal squamous cell carcinoma invading the muscularis mucosae or deeper. Gastrointest Endosc. 2002 Sep;56(3):387–90.

[7]. Trivedi PJ, Braden B. Indications, stains and techniques in chromoendoscopy. QJM. 2013 Feb;106(2):117–31.

[8]. Nagami Y, Tominaga K, Machida H, Nakatani M, Kameda N, Sugimori S, et al. Usefulness of non-magnifying narrow-band imaging in screening of early esophageal squamous cell carcinoma: a prospective comparative study using propensity score matching. Am J Gastroenterol. Nature Publishing Group; 2014, 109(6):845–54.

[9]. van der Sommen F, Zinger S, Curvers WL, Bisschops R, Pech O, Weusten BLAM, et al. Computer-aided detection of early neoplastic lesions in Barrett's esophagus. Endoscopy. 2016 Apr 21.

[10]. Everson M, Herrera LCGP, Li W, Luengo IM, Ahmad O, Banks M, et al. Artificial intelligence for the real-time classification of intrapapillary capillary loop patterns in the endoscopic diagnosis of early oesophageal squamous cell carcinoma: A proof-of-concept study. United European Gastroenterology Journal. 2019, 7(2):297-306.

[11]. Girshick R., Donahue J., Darrell T., and Malik J., Rich feature hierarchies for accurate object detection and semantic segmentation, In Proc. IEEE Conf. on computer vision and pattern recognition (CVPR), pp. 580-587. 2014.

[12]. Uijlings, J., van de Sande, K., Gevers, T., & Smeulders, A. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*, 154–171.

[13]. Girshick R., Fast R-CNN, In Proc. IEEE Intl. Conf. on computer vision, pp. 1440-1448. 2015.

[14]. S. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks, In Advances in neural information processing systems (NIPS), pp. 91-99. 2015.

[15]. Redmon J., Divvala S., Girshick R., and Farhadi A., You only look once: Unified, real-time object detection, CVPR 2016.

[16]. Redmon J., Farhadi A., YOLO9000: Better, Faster, Stronger, CVPR 2017.

[17]. Redmon J., Farhadi A., YOLOv3: An incremental improvement, arXiv: 1804.02767, arXiv.org, 2018.

[18]. Liu W., Anguelov D., Erhan D., Szegedy C., and Reed S., SSD:Single shot multibox detector,arXiv:1512.02325, 2015.

[19]. Lin T.-Y., Goyal P., Girshick R., He K., and Doll´ar P., Focal loss for dense object detection.arXiv preprintarXiv:1708.02002, 2017

[20]. K. He, G. Gkioxari, P. Doll´ar, and R. Girshick. Mask R-CNN, arXiv:1703.06870, 2017.

[21]. Mask-RCNN, https://github.com/matterport/Mask_RCNN.

[22]. YOLOv3, https://github.com/qqwweee/keras-yolo3.

[23]. COCO dataset: http://cocodataset.org/#home.

[24]. Rosebrock, A., Non-Maximum Suppression for Object Detection in Python, https://www.pyimagesearch.com/2015/02/16/faster-non-maximum-suppression-python/ .

[25]. VIA, http://www.robots.ox.ac.uk/~vgg/software/via/.

[26]. Gao X., Y. Wang, Y. Qian, A. Gao, Modelling of chromatic contrast for retrieval of wallpaper images, Color Research and Application, 40(4):361-373, 2015.